

Depression Detection in Tweets using Logistic Regression Model

Rahul Kumar Sharma¹, Vijayakumar A²

¹Final Year MCA Student, Dept of MCA, School of CS & IT, Jain Deemed-to-be University, Bengaluru

²Professor, Dept of MCA, School of CS & IT, Jain Deemed-to-be University, Bengaluru

Email: rahulsharma.rks31@gmail.com¹, vijay.pattukkottai@gmail.com²

ABSTRACT

In the growing world of modernization, mental health issues like depression, anxiety and stress are very normal among people and social media like Facebook, Instagram and Twitter have boosted the growth of such mental health. Everything has its legitimacy and negative mark. During this pandemic, people are more likely to suffer from mental health issues, they are available 24*7 and are cut off from the real world. Past examinations have shown that individuals who invest more energy via online media are bound to be depressed. In this project, we find out people who are depressed based on their tweets, followers, following and many other factors. For this, I have trained and tested our text classifier, which will distinguish between the user who is depressed or not depressed.

KEYWORDS: Depression, Flask, Mental Health, Natural Language Toolkit (NLTK), Twitter, Wordcloud

How to cite this paper: Rahul Kumar Sharma | Vijayakumar A "Depression Detection in Tweets using Logistic Regression Model"

Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-4, June 2021, pp.724-727, URL: www.ijtsrd.com/papers/ijtsrd41284.pdf



IJTSRD41284

Copyright © 2021 by author (s) and International Journal of Trend in Scientific Research and Development Journal. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0) (<http://creativecommons.org/licenses/by/4.0>)



I. INTRODUCTION

Mental health like depression is most common in people these days. Ten in every hundred people suffer from mental health illness. This issue influences the individual perspective, mind-set, feeling and conduct with adjoining individuals. Most of the time people don't talk about their mental health which sometimes leads to severe health issues. Depression is one of the primary reasons that cause disability worldwide. In the pre-mature age of depression, 70% of people don't consult with a doctor or with anyone which can cause serious damage in future. There is a movement going on to leverage social media data for finding, estimating, and tracking the changes in the occurrence of disease. The presence of online media gives us the freedom to improve the information accessible to psychiatrist and researchers enabling them to be better informed. Depression is viewed as the superb justification individuals who will in general have self-destructive contemplations. Around 80% who attempt or die by suicide are suffering from depression. But because of denial or ignorance, it is undetected and people suffer. It can be prevented if it is diagnosed in its early stage. This project will help us to find whether the person is depressed or not depressed based on their tweets. It can be used to find signs of other issues related to mental health. In this study, we used tweets from different users, which are labelled as 0 and 1. If the result is 0, it means the user is not depressed and if the result is 1, it means the user is depressed.

II. Literature Survey

In the past few years, researchers have worked to find ways to detect depression in tweets of user based on his/her tweets. The number of systems that were proposed to find depression are large in number. We can use sentiment analysis on text, web scrapping of tweets. Thousands of data collected using mining algorithm that can bring some meaningful insight into data.

Jamil Z (2017). proposed monitoring of tweets for depression detection to detect at-risk users. University of Ottawa [1]. SVM, Naive Bayes (NB), and Decision Tree (DT) are a portion of the generally utilized calculations in natural language processing tasks. SVM-linear classifier gives us the best result. As we all know no one algorithm can perform all task. Scientists attempt to alter the calculation as indicated by their, necessities of the task.

a. Decision Tree

Decision Tree classifies the instance based on the feature values. Han J., Pei J., & Kamber M. (Elsevier,2011) gave the concept and technique of Data mining [2]. Every node in the decision tree addresses an element and every division addresses a value that the node can undertake. A decision tree involves partitioning the data into subdivisions that contain occurrences with similar values, utilizing a significant perspective in a decision tree, called split choice, which means to discover a quality and its connected parting

capacity for each test node in a decision tree. Podgorelec, V., & Zorman, M. (2014) gave insight about Decision tree learning in Encyclopaedia of Complexity and Systems Science 1-28 [3]. Splits are evaluated by calculating entropy.

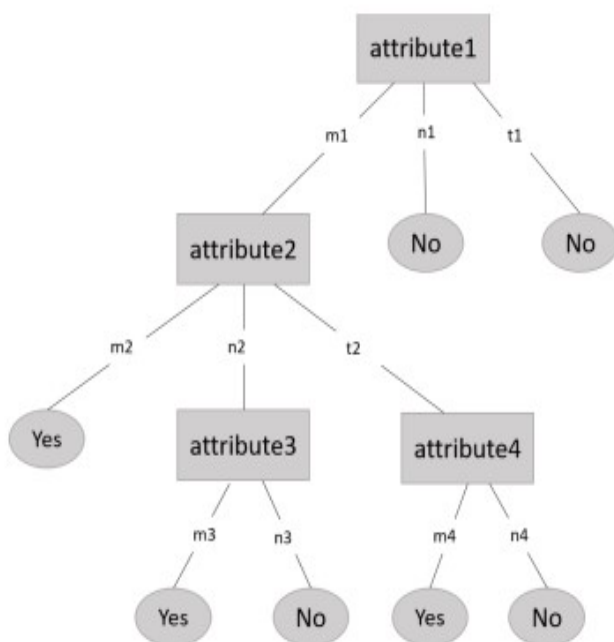


Fig:2.1 Decision Tree

b. Random Forest

The Random forest classifier makes different choice trees from an arbitrarily chosen subset of the preparation dataset. Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol Sanchez, J. P. (2012) gave an assessment for the effectiveness of a random forest classifier for landcover classification. ISPRS Journal of Photogrammetry and Remote Sensing 67: 93-104 [4]. At that point, it adds up to the votes from various choice trees to choose for the last class of test objects. Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A.R., Kundu, S. (2018) proposed an improved random forest for classification IEEE Transactions on Image Processing 27 (8): 4012-4024 [5]. A random forest classification was proposed in with a diminished number of trees.

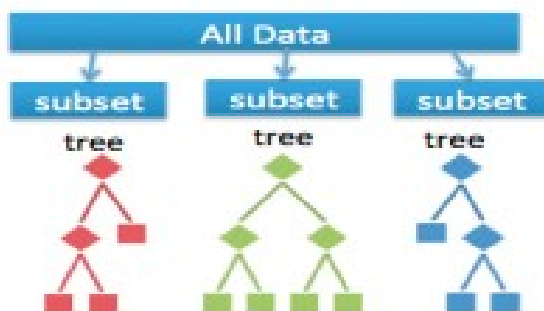


Fig:2.2 Random Forest

c. Support Vector Machine

Saitta L., (2000) proposed "Support Vector Networks." Kluwer Acad. Publ. Bost.: 273-297 [6]. A Support Vector Machine is a machine learning algorithm that works for both regression and grouping task, in any case, is mainly used in arrangement. Hamed, T., Dara, R., Kremer, S. C. (2014) gave an accurate, fast embedded feature selection for SVMs in 2014 13th International Conference on Machine Learning and Applications, IEEE :135-140 [7]. This classifier has been used of late in numerous applications because of its

outstanding characterizing capacity and show quality, isolating the information directly into two separate classes, with the most extreme distance between the two classes.

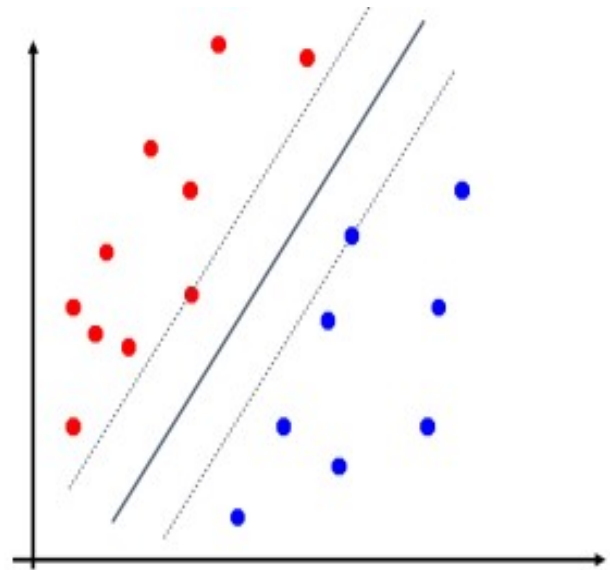


Fig:2.3 Support Vector Machine

d. K-Nearest Neighbour

K-Nearest Neighbour (K-NN) is potentially the clearest computations embraced in machine learning for classification and regression issues. In light of the nearest gauges, KNN takes data and arranges ongoing data focuses. The data is then assigned to the class with the premier nearest neighbour. Taneja, S., Gupta, C., Goyal, K., Gureja, D. (2014) proposed an enhanced k nearest neighbour algorithm using information gain and clustering in Fourth International Conference on Advanced Computing & Communication Technologies IEEE:325-329. [8] KNN is regularly used to characterise future data because of its simplicity of execution and sufficiency.

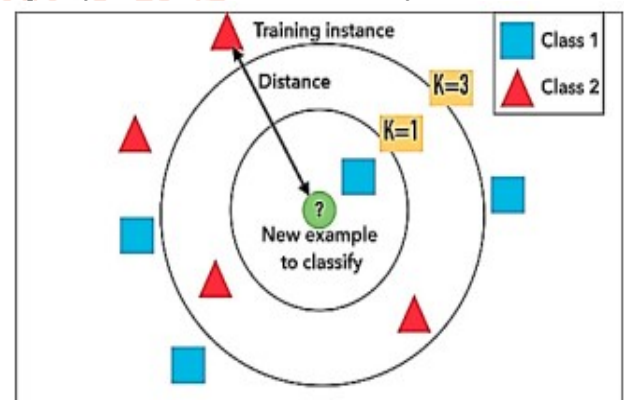


Fig:2.4 K-Nearest Neighbour

III. Existing System

In the developing universe of the web, individuals like to live in a virtual world. They share every thought-on platform like Facebook, Instagram, Twitter. They will in general contrast themselves as well as other people. Past examinations have shown that individuals who invest more time in online media are bound to be depressed. If depression is detected in an early stage, it can be cured, in later stages, it becomes more difficult to cure.

The project is concerned about that whether the person is depressed or not based on their tweet.

IV. Proposed System

The goal of this project is to build a model which can help us to detect whether the person is suffering from depression or not based on their tweet. People can compose a tweet in the content box and it will be breaking down by the model that we have made and it will give us the outcome. I will be using the python web framework Flask to integrate it with the model and make it more intractable to the common user.

The proposed system can help us to make people aware of their mental health and they can take necessary measure and help themselves. We take the dataset and clean it with the justifiable goal. The dataset contains tweets and label (0 and 1). If the tweet is depressing then the result will be 1 and if the tweet is not depressing the result will be 0. I will be using the Logistic Regression Model and check the accuracy. Later the model will be saved using a pickle library.

a. Logistic Regression Model

In statistics, the logical model is utilized to demonstrate the likelihood of a specific class or occasion existing like win/lose, alive/dead or healthy/unhealthy. Logistic Regression is a characterization algorithm. It is utilized to foresee a binary outcome based on a set of independent variables. A binary result is one where there are just two potential situations either the occasion occurs (1) or it doesn't occur (0). Independent factors are those factors or factors which may impact the result (or ward variable).

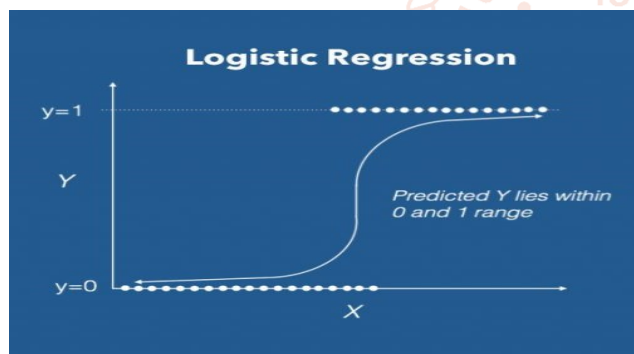


Fig:4.1 Logistic Regression

b. Proposed System Architecture

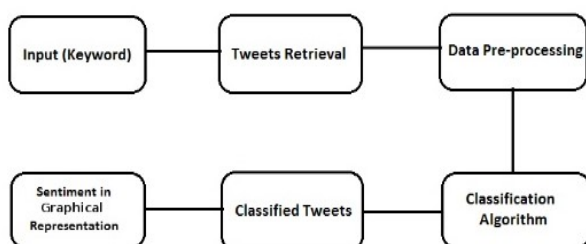


Fig:4.2 Proposed System Architecture Twitter Sentiment Analysis

As we can see in the above diagram, the user logs in through their twitter account, and writes tweet. Text pre-processing is applied to the tweet. Corpus is created, tweet is tokenized and in last normalization is applied, where all character is converted to lower case, links, emoji, punctuations are removed. Hornik, K., & Grün, B. Topicmodels: used an R package for fitting topic models. J. of Stat. Softw. 40 (13), 1-30 (2011) [9]. Stemming is applied and a document term network (DTM) is made for each record. The matrix shows

the reoccurrence of words in each tweet, where each line demonstrates an archive of tweets and every segment demonstrates all words utilized in all records. TF-IDF is used to measure the words' weight. Features are applied on to the DTM are then converged with account measures removed from the social network and user activities. Result of the merge are then treated as free factors in classification algorithms to anticipate the reliant variable of a result of interest. Ultimately, we decide upon the Logistic Regression algorithm.

V. Result and Discussions

	precision	recall	f1-score	support
0	0.96	0.99	0.98	2011
1	0.99	0.96	0.98	2011
accuracy			0.98	4022
macro avg	0.98	0.98	0.98	4022
weighted avg	0.98	0.98	0.98	4022

Fig:5.1 Result of the Logistic Regression Model

By taking a look at the above picture, we can see that the model has performed well. The exactness of our model is 96%. Now we can use this model to integrate it into our template so that it can be used by the users.

a. Model Comparison

Table 1: Accuracy and Model comparison table

Model	Accuracy	Precision	F1-measure
Decision Tree	72.5	0.473684	0.5625
Support Vector Machine	72.5	0.681818	0.652174
Logistic Regression	98	0.98	0.98

In order to get better result, we are going to compare our model with other models which have different accuracy. In the above picture we can see that the Logistic Regression have the highest accuracy.

VI. Conclusion and Future Enhancements

Machine Learning and Artificial Intelligence assume a significant part in Healthcare, Banking, Stocks, Cyber Security, Weather Forecast, etc. We as a client gather information, clean it, train it, test it and make a forecast. The exactness of the outcome relies upon the nature of the information and the type of algorithm used. Many pre-processing steps are performed, including data preparation and aligning, data labelling, and feature extraction and selection. The Logistic Regression Model has accomplished its ideal exactness and this model can be utilized to foresee that if the client is depressed or not. This project can be considered a step forward to build a social media platform for analyzing user activity and to predict the state of mental health of the user.

Utilizing this stage, the client doesn't need to make an account. They can compose any content or tweet, and the model will investigate if the individual is depressed or not. With the help of this result, the user can take precautions and can consult with the psychiatrist. It can act as a standalone application for depression detection. In future, we can use voice and facial expression to detect depression.

Reference-

- [1] Jamil, Z. Monitoring tweets for depression to detect at-risk users. Université d'Ottawa/University of Ottawa, 2017.
- [2] Han J., Pei J., & Kamber, M. Data mining: concepts and techniques (Elsevier,2011)
- [3] Podgorelec, V., & Zorman, M. Decision tree learning in Encyclopedia of Complexity and Systems Science 1-28 (2014)
- [4] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol Sanchez, J. P. (2012) "An assessment of the effectiveness of a random forest classifier for landcover classification." ISPRS Journal of Photogrammetry and Remote Sensing 67: 93-104.
- [5] Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintla, A.R., Kundu, S. (2018)"Improved random forest for classification."IEEE Transactions on Image Processing 27 (8): 4012-4024.
- [6] Saitta, L., (2000) "Support Vector Networks." Kluwer Acad. Publ. Bost.: 273-297.
- [7] Hamed, T., Dara, R., Kremer, S. C. (2014) "An accurate, fast embedded feature selection for SVMs." In2014 13th International Conference on Machine Learning and ApplicationsIEEE :135-140.
- [8] Taneja, S., Gupta, C., Goyal, K., Gureja, D. (2014) "An enhanced k nearest neighbour algorithm using information gain and clustering." Fourth International Conference on Advanced Computing & Communication Technologies IEEE:325-329.
- [9] Hornik, K., & Grün, B. Topicmodels: An R package for fitting topic models. J. of Stat. Softw. 40 (13), 1-30 (2011).

